APPLICATION OF

# 迁移学习：理论与实践

COMPACT CODING IN

邵　浩◎著

TRANSFER LEARNING

上海交通大學出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

本书由教育部人文社会科学研究青年基金项目(编号:13YJC630126),上海对外经贸大学 085 工程重点学科专业建设项目(编号:Z085FX13003)资助

# Applications of Compact Coding in Transfer Learning

# 迁移学习:理论与实践

邵 浩 著

上海交通大学出版社

## 内容提要

本书着眼于管理实际中的资源再利用，对数据挖掘领域最前沿的迁移学习(transfer learning)进行了详细阐述，并着重介绍了应用最为广泛的分类学习(classification)，将最前沿的研究进行了归纳总结，并通过实际算法分析，将领域内的最新进展提供给读者，使读者能够使用迁移学习的工具构建模型并应用到实际问题。本书主要读者对象为具有管理和计算机背景并在数据挖掘领域有初步研究的学者。

**图书在版编目(CIP)数据**

迁移学习:理论与实践/邵浩著. —上海:上海交通大学出版社,2013
ISBN 978 - 7 - 313 - 10656 - 8

Ⅰ. ①迁… Ⅱ. ①邵… Ⅲ. ①数据采集—研究 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2013)第 283507 号

**迁移学习:理论与实践**

著　　者: 邵　浩
出版发行: 上海交通大学出版社　　地　　址: 上海市番禺路 951 号
邮政编码: 200030　　电　　话: 021 - 64071208
出 版 人: 韩建民
印　　制: 常熟市梅李印刷有限公司　　经　　销: 全国新华书店
开　　本: 787mm×960mm　1/16　　印　　张: 8.25
字　　数: 120 千字
版　　次: 2013 年 12 月第 1 版　　印　　次: 2013 年 12 月第 1 次印刷
书　　号: ISBN 978 - 7 - 313 - 10656 - 8/TP
定　　价: 29.00 元

# Preface

Nowadays, classification technologies in data mining and machine learning are widely implemented to various applications such as disease diagnoses and biological taxonomy. However, we are often confronted with the lack of labeled data for a new task especially in medical or biological data sets. For example, in the disease examination, with only a few diagnosed cases, it is difficult to build an accurate model for more cases in the future. Transfer learning and active learning are two separate strategies to compensate for the insufficiency of the labeled data. The former approach builds novel models on new tasks by extracting useful knowledge from the existing ones, and the latter approach is designed to select informative unlabeled instances and consult experts for the labels. Although it is natural to consider borrowing knowledge from similar tasks, despite numerous researches on transfer learning, few of them have a solid theoretical framework and are parameter-free. Furthermore, how to avoid the negative transfer problem, which happens due to different distributions between tasks is still an open question. In active learning, the initial hypothesis re-sampled from limited labeled instances is often inaccurate and unreliable.

As a new research area, to the best of our knowledge, there exists no systematic book for the introduction of transfer learning and its applications. In this book, we aim to introduce the background of transfer learning, as well as some typical algorithms, to shed lights for researchers who are new to this research topic on the methodologies for solving the transfer learning problems. We lay emphases on several representative algorithms in different chapters. We will discuss transfer learning algorithms with the Minimum Description Length Principle (MDLP), which is based on a solid theoretical foundation and free from parameters. The basic MDLP is extended by introducing the code book to build a connection between the source and the target domains. A compact coding method for hyperplane classifiers in inductive transfer learning setting is also introduced. In the framework, a two-level evaluation is proposed to measure the similarities between different tasks, which are represented by the code lengths through minimum encoding. In such a way, the weights of dissimilar source tasks are decreased iteratively to avoid negative transfer on the target task. Then we introduce an algorithm to improve active learning using the transferred knowledge. This algorithm is designed by extending the usage of a basic active learning scenario with a divergence measure, in order to select only informative unlabeled instances for query and obtain a high classification accuracy. Meanwhile, an adaptive strategy is designed to eliminate unnecessary instances and inferior models. Furthermore, some classical algorithms such as Gaussian Process are incorporated in the transfer learning scenario.

For each algorithm, we perform extensive experiments on both synthetic and real data sets to test the effectiveness of the proposed algorithms, The classification accuracy is generally improved by 5% to 10% compared to the state-of-the-art methods. It is also proved that these algorithms are robust against noise. In active learning scenario, the algorithm is proved to be able to select less queries with a high accuracy than other methods.

# Contents

# Chapter 1

# Introduction

## 1.1 Background and Motivation

Despite the significant achievements on the development of machine learning and data mining technologies, most of them make the identical assumption that the training data and the test data are under the same distribution. However, in real applications, we are often confronted with problems without or with only a few labeled instances in the target task. Therefore the training data with the same distribution is far from sufficient to obtain a satisfactory result by traditional methods. For example, a challenging issue in nosology is that when a new epidemic strikes, it is crucial to classify the patients as early as possible with a high accuracy, in order to provide effective treatments. However, even though there exist thousands of suspected cases, only a few of them are labeled (diagnosed). In such circumstances, we need to find effective ways by taking advantages of other sources to achieve a high classification accuracy.

Instead of learning a new task from scratch, we can naturally consider two methods, either to borrow existing knowledge from other similar data sets, or consulting experts/teachers to obtain supervised knowledge. In the example above, we may refer to useful models or instances from similar disease data sets, or select informative instances by asking doctors for authentic diagnoses to enlarge the number of labeled instances. In machine learning and data mining literatures, we call the two distinctive methodologies Transfer Learning (TL) and Active Learning (AL), respectively. They are both countermeasures against the problem of insufficient labeled instances in the target task.

According to S. J. Pan and Q. Yang, transfer learning is endowed with the ability to recognize and apply knowledge and skills learnt in the previous tasks to the new task in the new domain [53]. It may be classified into four categories in terms of the availability of labeled data in the source and the target learning tasks. The first case is supervised inductive transfer learning, where both the source and the target tasks contain labeled data. Multi-task learning belongs to this category where several tasks are learnt simultaneously [47, 7, 52]. The second case is self-taught learning with labeled data available in the target task but not in the source task [49]. The third case is transductive transfer learning in which labeled data are available in the source task but not in the target task, including domain adaptation, sample selection bias/covariate shift, etc. [25, 30, 12]. The last case is unsupervised transfer learning [58, 69], where neither the source task nor the target task contains labeled data, so that the available information to the learner is the least. We explicitly consider the learning scenario of supervised inductive transfer learning, in which a bundant labeled information is available in the source domain, which is probably helpful for learning novel models on new tasks. As shown in Figure 1.1, the target task is assumed to have only a few labeled instances. We can obtain satisfactory results by extracting useful knowledge from the existing models in the source domain, instead of applying learning methods once again, which is often expensive and time-consuming.

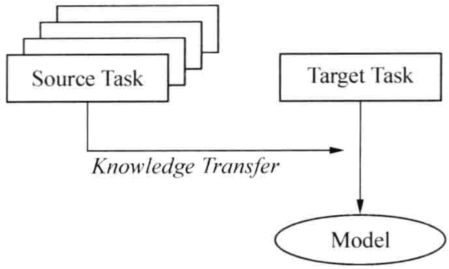


Figure 1.1: Illustration of Transfer Learning

While transferring, we need to pay attention to the problem that the distributions underlying the source and the target domains are not identical. Consider a typical example which is language study, where learning English is the target task, and learning Japanese and French is the source tasks. It would be much effective for a learner to study English if he (she) is familiar with French. However, the learning process would be impeded if the learner only masters Japanese, because the lexical and semantic structure of Japanese is different from that of English. When the distributions of the source and the target domains are different, directly transferring knowledge would hurt the performance on the target task, which is also known as "negative transfer" [41]. It is likely to happen once we underestimate the side effect resulting from the distribution differences of multiple source tasks, which is common in real applications. Moreover, to the best of our knowledge, none of the existing inductive transfer learning methods is free from parameters. It is widely known that a method with many parameters is subject to overfitting and the user has to tune parameters prior to learning [20]. Therefore, we are motivated to develop a method under a solid theoretical foundation and without parameters. The Minimum

Description Length Principle (MDLP) possesses the desired properties, it is a principle where the best hypothesis that can be inferred from data is the one with the shortest sum of the code length of the hypothesis and the code length of the data using the hypothesis. It has a solid theoretical framework and a clear interpretation, is robust to noise and requires no parameter specification. In our work, MDLP is extended to accommodate to the transfer learning scenario.

Active learning (AL) [9] provides a solution by selecting unlabeled instances to query, in order to obtain a satisfactory classifier using as few instances as possible where the labeling cost is high, as shown in Figure 1.2. It can be divided into two main categories, which are the stream-based active learning [19] and the pool-based active learning [31]. The former one scans the data sequentially and makes query decisions individually, while the pool-based active learning makes one decision each time from the entire collection of the unlabeled pool. For the pool-based active learning we firstly select an instance in the large unlabeled pool and ask an oracle(e. g., human annotator) to label it, then add the instance to the labeled training set. This process can iterate several times in order to achieve satisfactory results. A significant challenge is how to select the most informative instance in each query from the large amount of unlabeled instances in the pool in the target task. There has been substantial work on active learning, for example, the Query by Committee (QBC) model which assumes a correct Bayesian prior on the set of hypotheses, and the committee members are all trained on the current labeled set [27]. However, as pointed out in [40], given only a few labeled instances in the data set, the initial hypotheses sampled from these instances are not reliable since they may deviate from the optimal hypothesis with respect to the input distribution at the end of the classification procedure. Moreover, existing AL methods commonly assume that there exists a perfect hypothesis underlying the data set,

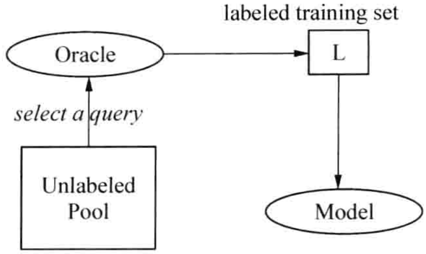


Figure 1.2: Illustration of Active Learning

which is not true in real applications where the distributions are usually unknown.

Recently, a new research area named as Transfer Active Learning (TAL) has drawn much attention. In this learning setting, transfer learning and active learning are adaptively integrated to compensate for the disadvantages of both methods, with the objective to obtain a high accuracy and to query as few samples as possible [26, 60]. However, to the best of our knowledge, no research has been proposed which concern both the negative transfer problem in transfer learning and the initial hypothesis problem in active learning. Moreover, there exists no effective strategy in these methods to avoid querying unnecessary instances which may increase the querying cost. We intend to design an adaptive active learning algorithm by adopting transfer learning techniques in order to obtain high classification accuracy with less quiresand avoid negative transfer at the same time.

## 1.2 Contributions

### 1.2.1 Extended MDLP for Transfer Learning

Transfer learning provides a solution in real applications of how to learn a target task where a large amount of auxiliary data from the source domains are given. Despite numerous studies on this topic, few of them have a solid theoretical framework and are parameter-free. We propose an Extended Minimum Description Length Principle (EMDLP) for feature-based inductive transfer learning, in which both the source and the target data sets contain class labels and relevant features are transferred from the source domain to the target one. Unlike conventional methods, our encoding measure is based on a theoretical back-ground and has no parameter. To obtain useful features to be used in the target task, we design an enhanced encoding length by adopting a code book which stores useful information obtained from the source task. With the code book which builds connections between the source and the target tasks, our EMDLP is able to evaluate the inferiority of the results of transfer learning with the add-sum of the code lengths of five components: those of the corresponding two hypotheses, the two data sets with the help of the hypotheses, and the set of the transferred features. The proposed method inherits the nice property of the MDLP that elaborately evaluates the hypotheses and balances the simplicity of the hypotheses and the goodness-of-the-fit to the data. Extensive experiments using both synthetic and real data sets show that the proposed method provides a better performance in terms of classification accuracy and is robust against noise.

### 1.2.2 Compact Coding for Hyperplane Classifiers in Transfer Learning

In real applications, a new task is often related to another existing task. Transfer learning techniques are developed to build novel models on new tasks by extracting useful information from the existing models, to reduce the high cost of inquiring the labeled information for the target task. However, how to avoid *negative transfer* which happens due to different distributions of tasks is still an open problem. Unlike traditional methods which only measure either similarity between tasks or instance relatedness, we propose a Compact Coding method for Hyperplane Classifiers (CCHC) under a *two-level* framework in inductive transfer learning setting. In the *macro level*, the degree of the similarity is represented by the relevant code length of the class boundary of each source task with respect to the target task through minimum encoding. In addition, informative instances of the source tasks are adaptively selected in the *micro level* to make the choice of the specific source task more accurate. Extensive experiments show the effectiveness of our algorithm in terms of the classification accuracy in both UCI and text data sets.

### 1.2.3 Transfer Active Learning

In real applications of inductive learning, labeled instances are often deficient and active learning plays a role as a countermeasure. However, an inevitable problem in active learning is that the initial hypotheses sampled from a few labeled instance are not reliable and may deviate from the true distribution underlying the target task. Consequently, a promising research is to compensate for the disadvantages of active learning with the help of transfer learning by borrowing the abundant labeled information from related data. A few studies have been done to integrate the two methods together, which is called transfer active learning. However, when there exists unrelated domains which have different distributions or label assignments, direct transfer will lead to degenerated performance. Although efforts are made to alleviate this problem in transfer learning by evaluating the similarities between tasks, due to the different problem setting in active transfer learning, there exists no effective strategy in the literature to avoid selecting unconcerned samples to query. To tackle these issues, we propose a hybrid algorithm for active learning with the help of transfer learning in which we explore the advantages of the KL divergence proved to be useful in transfer learning. We extend the usage of a basic active learning scenario to a more general setting in transfer learning, in which the hypotheses from the source domain are taken into consideration, and the KL divergence is adopted as the similarity between them, to decrease the negative effects of inferior hypotheses. To avoid querying unimportant instances, we also present an adaptive strategy which could eliminate unnecessary instances in the input space and models in the model space. Extensive experiments on both synthetic and real data sets show that our algorithm is able to query less instances and converges faster than the state-of-the-art methods.

### 1.2.4 Gaussian Process for Transfer Learning

To solve the problem of the deficiency of labeled instances, transfer learning techniques are introduced to make use of existing knowledge from the source data sets to the target data set. However, due to the discrepancy of distributions between tasks, directly transferring knowledge will possibly lead to degenerated performance which is also called *negative trasnfer*, which makes the classification of the target task difficult. We adopted the Gaussian process to alleviate this problem by directly evaluating the distribution differences, with the parameter-free Minimum Description Length Principle (MDLP) for encoding. The proposed method inherits the good property of solid theoretical foundation as well as noise-tolerance. Extensive experiments results show the effectiveness of our method.

## 1.3 Book Overview

In the next chapter, we start with a review of the relevant literature. Chapter 3 discusses our work for transfer learning based on the minimum description length principle. Chapter 4 describes our algorithm for encoding hyperplane classifiers in transfer learning. In Chapter 5, we develop an approach which combines active learning and transfer learning together in order to improve the classification accuracy. In Chapter 6 the Gaussian Process method for transfer learning is presented. Finally Chapter 7 gives the conclusion and the future research directions.